

# Context-Aware Scheduling of Joint Millimeter Wave and Microwave Resources for Dual-Mode Base Stations

Omid Semiari<sup>†</sup>, Walid Saad<sup>†</sup>, and Mehdi Bennis<sup>‡</sup>

<sup>†</sup>Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA,  
Emails: {osemiari, walids}@vt.edu

<sup>‡</sup> Centre for Wireless Communications, University of Oulu, Finland, Email: bennis@ee.oulu.fi

**Abstract**—One of the most promising approaches to overcome the drastic channel variations of millimeter wave (mmW) communications is to deploy dual-mode base stations that integrate both mmW and microwave ( $\mu$ W) frequencies. Reaping the benefits of a dual-mode operation requires scheduling mechanisms that can allocate resources efficiently and jointly at both frequency bands. In this paper, a novel resource allocation framework is proposed that exploits users' context, in terms of user application (UA) delay requirements, to maximize the quality-of-service (QoS) of a dual-mode base station. In particular, such a context-aware approach enables the network to dynamically schedule UAs, instead of users, thus providing more precise delay guarantees and a more efficient exploitation of the mmW resources. The scheduling of UAs is formulated as a one-to-many matching problem between UAs and resources and a novel algorithm is proposed to solve it. The proposed algorithm is shown to converge to a two-sided stable matching between UAs and network resources. Simulation results show that the proposed approach outperforms classical CSI-based scheduling in terms of the per UA QoS, yielding up to 36% improvement. The results also show that exploiting mmW resources provides significant traffic offloads reaching up to 43% from  $\mu$ W band.

## I. INTRODUCTION

Communication at high frequency, millimeter wave (mmW) bands is an effective way to boost the performance of 5G cellular networks [1], [2]. However, field measurements [2] have shown that the availability of mmW links can be highly intermittent, due to blockage by various obstacles. Therefore, meeting quality-of-service (QoS) constraints of delay-sensitive applications, such as HDTV and video conferencing, is challenging at mmW frequencies [2]–[8].

To provide a robust and reliable communication, mmW networks must coexist with small cell LTE networks that operate at the conventional microwave ( $\mu$ W) band [4]–[8]. In [4], the authors analyze transceiver architectures for dual-mode mmW- $\mu$ W networks. The work in [6] proposes a mmW- $\mu$ W dual-mode architecture used to transmit control and data signals, respectively, at  $\mu$ W and mmW frequency bands.

The problem of QoS provisioning for mmW is studied in [5], [7], and [8]. In [5], the authors propose a scheduling scheme that integrates device-to-device mmW links with 4G system to bypass the blocked mmW links. The work in [7] presented a mmW system at 60 GHz for supporting uncompressed high-definition (HD) videos for WLANs. In [8], the authors defined and evaluated important metrics to characterize multimedia QoS,

and designed a QoS-aware multimedia scheduling scheme to achieve the trade-off between performance and complexity.

Although interesting, the first body of work in [4]–[6] does not address the QoS provisioning in mmW- $\mu$ W networks. Moreover, [4], [7], and [8] do not consider multi-user scheduling and multiple access in dual-mode networks. In addition, conventional scheduling mechanisms, such as [5], [7], and [8], identify each user equipment (UE) by a single traffic stream with a certain QoS requirement. In practice, however, recent trends show that users run multiple applications simultaneously, each with a different QoS requirement. Even though the applications at a single device experience the same wireless channel, they may tolerate different delays and QoS, thus, resulting in different user's quality of experience. Accounting for precise, application-specific QoS metrics is particularly important for scheduling mmW resources whose channel is highly variable. In fact, conventional scheduling approaches fail to guarantee the QoS for multiple applications at a single UE.

With regard to QoS provisioning in dual-mode mmW- $\mu$ W networks, it is worthy to note that 1) traffic management mandates a joint scheduling that allocates resources in both frequency bands, and 2) the QoS constraint per user application (UA) can dictate whether the traffic should be served via mmW resources,  $\mu$ W resources, or both. Such knowledge of the *user's application context information*, is required for a robust and efficient scheduling in networks that incorporate dual-mode small cell base stations (SBSs).

The main contribution of this work is to propose a novel, context-aware resource allocation framework that intelligently allocates mmW and  $\mu$ W resources of a dual-mode SBS, depending on the specific delay constraints of UAs. This proposed context-aware scheduler allows each user to seamlessly run multiple applications simultaneously, each with certain QoS constraint. We formulate the problem as a two-sided matching game that aims to allocate time-frequency resources to UAs. To solve the game, we propose a novel distributed algorithm that allows UAs to submit requests for network resources based solely on their local information, i.e., tolerable delay and currently perceived network state. We show that the proposed algorithm yields a two-sided stable resource allocation to UAs. Simulation results show that the dual-band scheduler provides significant performance advantages, in terms of traffic offload, efficient mmW exploitation, and improved overall UA delay.

The rest of this paper is organized as follows. Section II presents the problem formulation. Section III presents the pro-

posed matching solution. Simulation results are analyzed in Section IV. Section V concludes the paper.

## II. SYSTEM MODEL

Consider the downlink of a dual-mode small base station (SBS) that operates at microwave ( $\mu\text{W}$ ) and millimeter wave (mmW) frequency bands. The coverage area of the SBS is a planar area with radius  $r$  centered at  $(0, 0) \in \mathbb{R}^2$ . Moreover, there are  $M$  UEs in the set  $\mathcal{M}$ , distributed randomly and uniformly within the SBS coverage. UEs are equipped with both mmW and  $\mu\text{W}$  RF interfaces which allow them to manage their traffic at both frequency bands. In addition, a UE at distance  $d$  from the SBS will experience a LoS mmW connection with probability  $\rho$  if  $d \leq r$ , otherwise,  $\rho = 0$ .

The antenna arrays of mmW transceivers allow to achieve an overall gain of  $\psi(x_1, x_2)$  for a UE located at  $(x_1, x_2) \in \mathbb{R}^2$  [9]. On the other hand, the transceivers at  $\mu\text{W}$  frequency have conventional single element omni-directional antennas. Furthermore, each UE  $m \in \mathcal{M}$  runs  $\kappa_m$  UAs. We let  $\mathcal{A}$  be the set of all UAs with  $A = \sum_{m \in \mathcal{M}} \kappa_m$  as the total number of UAs across all UEs.

### A. Multiple Access at mmW and $\mu\text{W}$ Frequency Bands

At mmW band, time division multiple access (TDMA) is used to schedule UAs at time slots of duration  $\tau_1$  [5]. At each mmW time slot, the SBS will transmit the bits associated with one UA over  $K_1$  resource blocks (RBs). To model large-scale channel effects at mmW links, we use the popular model of [9]:

$$L_1(x_1, x_2) = \beta_1 + \alpha_1 10 \log_{10}(\sqrt{x_1^2 + x_2^2}) + \chi, \quad (1)$$

where  $L_1(x_1, x_2)$  is the path loss at mmW frequencies for all UAs associated with a UE located at  $(x_1, x_2) \in \mathbb{R}^2$ . In fact, (1) is known to be the best linear fit to the propagation measurement in mmW frequency band [9], where  $\alpha_1$  is the slope of the fit and  $\beta_1$ , the intercept parameter, is the pathloss (in dB scale) for 1 meter of distance. In addition,  $\chi$  models the deviation in fitting (in dB scale) which is a Gaussian random variable with zero mean and variance  $\xi_1^2$ . Overall, the total achievable rate for UA  $a$  at time slot  $j$  is given by

$$R_a(j) = \sum_{k=1}^{K_1} w_1 \log_2 \left( 1 + \frac{p_{k,1} \psi(x_1, x_2) |h_{kj}|^2 10^{-\frac{L_1(x_1, x_2)}{10}}}{w_1 N_0} \right), \quad (2)$$

where  $w_1$  is the bandwidth of each RB,  $h_{kj}$  is the Rayleigh fading channel coefficient at RB  $k$  of slot  $j$ , and  $N_0$  is the noise power. The total transmit power at mmW band,  $P_1$ , is assumed to be distributed uniformly among all RBs, i.e.,  $p_{k,1} = P_1/K_1$ . For the  $\mu\text{W}$  band, we consider orthogonal frequency division multiple access (OFDMA) scheme in which multiple UAs can be scheduled over  $K_2$  RBs in the set  $\mathcal{K}_2$  at each  $\mu\text{W}$  time slot with duration  $\tau_2$ . Therefore, the achievable rate for an arbitrary UA  $a$  at RB  $k$  and time slot  $t$  is:

$$R_a(k, t) = w_2 \log_2 \left( 1 + \frac{p_{k,2} |g_{kt}|^2 10^{-\frac{L_2(x_1, x_2)}{10}}}{w_2 N_0} \right). \quad (3)$$

where  $w_2$  denotes the bandwidth of each RB at  $\mu\text{W}$  band, and  $g_{kt}$  is the Rayleigh fading channel at RB  $k$  at time-slot  $t$ . Moreover, the total transmit power at  $\mu\text{W}$  band,  $P_2$ , is assumed to be

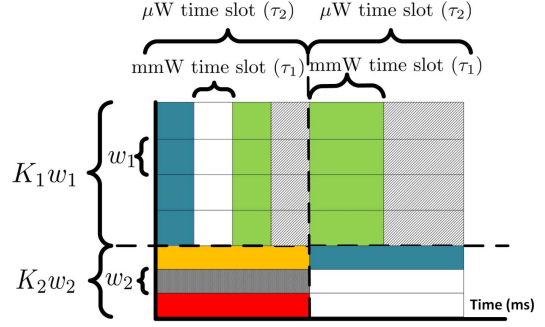


Fig. 1. Example of resource allocation of the dual-band configuration. Colors correspond to different UAs that may run at different UEs.

distributed uniformly among all RBs, i.e.,  $p_{k,2} = P_2/K_2$ . The path loss  $L_2(x_1, x_2)$  follows the log-distance model, similar to (1), with parameters  $\alpha_2$ ,  $\beta_2$ , and  $\xi_2$  adapted for  $\mu\text{W}$  band.

Hereinafter, unless otherwise specified, we use “time slot” to refer to the  $\mu\text{W}$  time slot. The scheduler allocates resources to UAs at the beginning of each time slot which remains unchanged for the next  $\tau_2$  seconds. Since mmW operates at very high frequencies, its channel coherence time will be relatively smaller than that of the  $\mu\text{W}$  frequencies [8]. Therefore, we let  $\tau_1 = \tau_2/J(t)$ , where  $J(t)$  is the number of UAs that will be scheduled in the mmW band at time slot  $t$ .

The proposed dual-band multiple access scheme is shown in Fig. 1, where each color identifies a single UA.

### B. Traffic Model

We assume a non-full buffer traffic model, where each UA has  $B$  bits of data to transmit. Once  $B$  bits are transmitted, the corresponding UA will be removed from the scheduling. Each UA has an application-specific tolerable delay which specifies its QoS class.

**Definition 1:** The QoS class,  $\mathcal{A}_t$ , is defined as the set of all UAs over all UEs that can tolerate a packet transmission delay of  $t$  time slots.

Our system has a total of  $T$  QoS classes,  $\bigcup_{t=1}^T \mathcal{A}_t = \mathcal{A}$ , and  $\mathcal{A}_t \cap \mathcal{A}_{t'} = \emptyset, t \neq t'$ . Due to system resource constraints, not all UAs can be served instantly, thus, access delay may occur to UAs. In fact, to transmit a data stream of size  $B$  bits to UA  $a \in \mathcal{A}_t$ , an average data rate of  $B/t\tau_2$  during  $t$  consecutive time slots is needed. The context information for all UAs is captured by the set  $\mathcal{C} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_T\}$ .

### C. Problem Formulation

At each time slot  $t$ , a scheduling decision  $\pi_t$  assigns time-frequency resources to UAs, at both mmW and  $\mu\text{W}$  bands. The scheduler takes the context information  $\mathcal{C}$  and achievable rates  $R_a(k, t)$  and  $R_a(j)$ , for all  $a \in \mathcal{A}$ ,  $k = 1, \dots, K_2$ , and  $j = 1, \dots, J(t)$ , as inputs and outputs integer variables  $x_{akt}$  and  $y_{ajt}$ .  $x_{akt} = 1$ , if RB  $k$  at time slot  $t$  is allocated to UA  $a$  and  $x_{akt} = 0$ , otherwise. In addition,  $y_{ajt} = 1$ , if mmW time slot  $j$  at slot  $t$  is allocated to UA  $a$ , otherwise  $y_{ajt} = 0$ .

The scheduling decision at a given slot  $t$  depends on scheduling decisions at previous time slots  $\{\pi_1, \pi_2, \dots, \pi_{t-1}\}$ . Thus, we define  $\pi = \{\pi_1, \pi_2, \dots, \pi_t, \dots, \pi_M\} \in \Pi$  as a scheduling policy, where  $\Pi$  is the set of all possible scheduling policies. For a given

policy  $\pi$ , the average rate for UA  $a$  after time slot  $t$ ,  $\bar{R}_\pi(a, t)$ , is

$$\bar{R}_\pi(a, t) = \frac{1}{t\tau_2} \sum_{t'=1}^t \left[ \tau_2 \sum_{k=1}^{K_2} R_a(k, t') x_{akt'} + \tau_1 \sum_{j=1}^{J(t')} R_a(j) y_{ajt'} \right]. \quad (4)$$

Next, we use (4) to formally define  $\mathbb{I}_a(\pi)$  as the QoS criterion for  $a \in \mathcal{A}_t$  as follow

$$\mathbb{I}_a(\pi) = \begin{cases} 0 & \text{if } \bar{R}_\pi(a, t) \geq \frac{B}{t\tau_2}, \\ 1 & \text{otherwise,} \end{cases} \quad (5)$$

where  $\mathbb{I}_a(\pi) = 0$  indicates that enough resources are allocated to UA  $a$ , to receive  $B$  bits within  $t$  slots.

Directional transmissions at mmW band compel the SBS to steer the beam toward one UE at any given mmW time slot [1]. Hence, without prior information of the channel for all UAs, the scheduler cannot schedule UAs based on the achievable rates,  $R_a(j)$ . Therefore, the scheduler follows an opportunistic round-robin (RR) scheme<sup>1</sup> at mmW band to allocate time slots to UAs. Therefore, the dual-mode SBS must be able to exploit the  $\mu$ W resources in order to meet the QoS constraints of UAs, i.e., to minimize  $\sum_{t=1}^T \sum_{a \in \mathcal{A}_t} \mathbb{I}_a(\pi)$ .

To this end, we focus on a subset of scheduling policies  $\pi \in \Pi^c \subseteq \Pi$  that schedules UAs in  $\mathcal{A}_t$  at mmW band for the first  $t-1$  time slots. Furthermore, the scheduler satisfies their required average rate,  $\bar{R}_\pi(a, t)$ , by efficiently allocating the resources of  $\mu$ W band at time slot  $t$  to UAs in  $\mathcal{A}_t$ . Hence, given  $\pi \in \Pi^c$ , we can write (4) as

$$\bar{R}_\pi(a, t) = \frac{1}{t\tau_2} \left[ \tau_2 \sum_{k=1}^{K_2} R_a(k, t) x_{akt} + B_\pi(a, t-1) \right], \quad (6)$$

where  $B_\pi(a, t-1)$  is the total number of transmitted bits for UA  $a$  up until time slot  $t$ , i.e.,

$$B_\pi(a, t-1) = \tau_1 \sum_{t'=1}^{t-1} \sum_{j=1}^{J(t')} R_a(j) y_{ajt'}. \quad (7)$$

Now, we formulate the problem as follows, for all  $t = 1, \dots, T$ :

$$\underset{\pi \in \Pi^c}{\text{minimize}} \quad \sum_{t=1}^T \sum_{a \in \mathcal{A}_t} \mathbb{I}_a(\pi), \quad (8a)$$

$$\text{s.t.} \quad \sum_{a=1}^A B_\pi(a, T) \leq B_{\text{tot}}, \quad (8b)$$

$$\sum_{j=1}^{J(t)} y_{ajt} \leq 1, \forall a \in \mathcal{A}, \quad (8c)$$

$$\sum_{a=1}^A y_{ajt} \leq 1, j = 1, \dots, J(t), \quad (8d)$$

$$\sum_{k=1}^{K_2} x_{akt} \leq K_2, \forall a \in \mathcal{A}, \quad (8e)$$

$$\sum_{a=1}^A x_{akt} \leq 1, k = 1, \dots, K_2, \quad (8f)$$

$$\sum_{k=1}^{K_2} \sum_{j=1}^{J(t)} x_{akt} y_{ajt} = 0, \forall a \in \mathcal{A}. \quad (8g)$$

In (8a), the objective is to maximize the QoS for UAs, using both mmW and  $\mu$ W resources. From (5), we can see that the objective function incorporates the context information  $\mathcal{C}$ . (8b) implies that the total transmitted bits after time slot  $T$ ,  $B_\pi(M)$ , must be less or equal to the total load  $B_{\text{tot}} = B \cdot A$ . (8c)-(8d) ensure orthogonal time-slot allocation for the mmW band. (8e)-(8f) guarantee orthogonal RB allocation at  $\mu$ W with OFDMA. Furthermore, (8g) implies that a single UA cannot be simultaneously assigned to both mmW and  $\mu$ W bands. Next, we propose a framework to solve the optimization problem.

### III. CONTEXT-AWARE SCHEDULING AS A MATCHING GAME

In (6) and (8a), we observe that  $B_\pi(a, t-1)$  is the only required information from previous time slots to schedule UA  $a \in \mathcal{A}_t$  at time slot  $t$ . Thus, at each time slot  $t$  we have:

**Remark 1:** The optimal scheduling decision  $\pi_t^*$  at time slot  $t$  is the solution of

$$\underset{\pi_t \in \Pi^c}{\text{maximize}} \quad \sum_{a \in \mathcal{A}_t} \sum_{k=1}^{K_2} R_a(k, t) x_{akt} \quad (9a)$$

$$\text{s.t.} \quad \sum_{k=1}^{K_2} R_a(k, t) x_{akt} \leq B - B_\pi(a, t-1), \quad \forall a \in \mathcal{A}_t, \quad (9b)$$

$$(8b) - (8g). \quad (9c)$$

The downlink scheduling problem in (9a)-(9c) is a combinatorial problem of matching users to resources which does not admit a closed-form solution and has an exponential complexity [10].

#### A. Scheduling as a Matching Game: Preliminaries

To solve the resource allocation problem in (9a)-(9c), we propose a novel solution based on matching theory, a mathematical framework that provides a decentralized solution with tractable complexity for combinatorial problems, such as the one in (9a)-(9c) [11], [12]. A *matching game* is defined as a two-sided assignment problem between two disjoint sets of players in which the players of each set are interested to be matched to the players of the other set, according to *preference relations*. At each time slot  $t$  of our scheduling problem,  $\mathcal{K}_2$  and  $\mathcal{A}_t$  are the two sets of players. A preference relation  $\succ$  is defined as a complete, reflexive, and transitive binary relation between the elements of a given set. Here, we let  $\succ_a$  be the preference relation of UA  $a$  and denote  $k \succ_a k'$ , if player  $a$  prefers RB  $k$  more than  $k'$ . Similarly, we use  $\succ_k$  to denote the preference relation of RB  $k \in \mathcal{K}_2$ .

In the proposed scheduling problem, the preference relations of UAs depend on both the rate and the QoS constraint. Matching theory allows to specify preference relations, by defining individual utility functions for UAs and SBS resources. In our scheduling game, the SBS will naturally control the preferences of all resources.

#### B. Dual-mode Scheduling as a Matching Game

Each scheduling decision  $\pi_t$  determines the allocation of RBs to UAs at time slot  $t$ . Thus, the scheduling problem can be defined as a *one-to-many matching game*:

<sup>1</sup> Other random access schemes can be readily accommodated in our model.



**Definition 2:** Given two disjoint finite sets of players  $\mathcal{A}_t$  and  $\mathcal{K}_2$ , the scheduling decision at time slot  $t$ ,  $\pi_t$ , can be defined as a *matching relation*,  $\mu_t : \mathcal{A}_t \rightarrow \mathcal{K}_2$  that satisfies 1)  $\forall a \in \mathcal{A}_t, \mu_t(a) \subseteq \mathcal{K}_2$ , 2)  $\forall k \in \mathcal{K}_2, \mu_t(k) \in \mathcal{A}_t$ , and 3)  $\mu_t(k) = a$ , if and only if  $k \in \mu_t(a)$ .

In fact,  $\mu_t(k) = a$  implies that  $x_{akt} = 1$ , otherwise  $x_{akt} = 0$ . Therefore,  $\mu_t$  can be viewed as a scheduling decision  $\pi_t$  that determines the allocation at  $\mu$ W band. One can easily see from the above definition that the proposed matching game inherently satisfies the constraints in (8e)-(8f). Next, we need to define suitable utility functions to determine the preference profiles of UAs and RBs. Given matching  $\mu_t$ , we define the utility of UA  $a$  for  $k \in \mathcal{K}_2$  at time slot  $t$  as:

$$\Psi_a(k, t; \mu_t) = \begin{cases} 0 & \text{if } a \notin \mathcal{A}_t \text{ or} \\ & \sum_{k' \in \mu_t(a)} R_a(k', t) \tau_2 > B - B_\pi(a, t' - 1), \\ R_a(k, t) & \text{otherwise.} \end{cases} \quad (10)$$

The utility of  $\mu$ W RBs  $k \in \mathcal{K}_2$  for UA  $a \in \mathcal{A}_t$  is simply the rate:

$$\Phi_k(a, t) = R_a(k, t). \quad (11)$$

Using these utilities, the preference relations of UAs and RBs at a given time slot  $t$  are:

$$k \succ_a k' \Leftrightarrow \Psi_a(k, t; \mu_t) \geq \Psi_a(k', t; \mu_t) \quad (12)$$

$$a \succ_k a' \Leftrightarrow \Phi_k(a, t) \geq \Phi_k(a', t), \quad (13)$$

for  $\forall a, a' \in \mathcal{A}$ , and  $\forall k, k' \in \mathcal{K}_2$ . We note that (10) and (12) depend on the context information  $\mathcal{C} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_T\}$ , while (11) and (13) rely only on the channel state information. Thus, the SBS will not need to know the delay tolerance of each UA, making the matching game suitable for distributed implementations.

### C. Proposed Context-aware Scheduling Algorithm

To solve the proposed game and find a suitable outcome, we use the concept of two-sided *stable matching* between UAs and RBs, defined as follows [11]:

**Definition 3:** A pair  $(a, k) \notin \mu_t$  is said to be a *blocking pair* of the matching  $\mu_t$ , if and only if  $a \succ_k \mu_t(k)$  and  $k \succ_a \mu_t(a)$ . Matching  $\mu_t$  is *stable*, if there is no blocking pair.

Under a stable matching  $\mu_t$ , one can ensure that the scheduler will not reallocate the RBs to other UAs. In fact, stability is a key requirement for distributed scheduling to ensure that UAs will not deviate from the solution that guarantees the QoS.

For conventional matching problems, the popular *deferred acceptance (DA)* algorithm is used to find a stable matching [11], [12]. However, DA cannot be applied directly to our problem because it assumes that the quota for each UA is fixed. The quota is the maximum number of RBs that a UA can be matched to. In our problem, however, quotas cannot be predetermined, since the number of required RBs to satisfy the QoS constraint of a UA, in (9b), depends on the channel quality at each RB.

In fact, the adopted utility functions in (10) depend on the current state of the matching. Due to the dependency of UAs' preferences to the state of the matching, i.e.  $x_{akt}$  variables,

TABLE I  
PROPOSED CONTEXT-AWARE SCHEDULING ALGORITHM

---

<b>Inputs:</b> $\mathcal{C}, \mathcal{A}, \mathcal{K}_2, B$ .	
<i>Initialize:</i> $t = 1$ ; $B_\pi(a, 0) = 0$ ; $\mathcal{K}_a = \mathcal{K}_2, \forall a \in \mathcal{A}$ .	
<b>while</b> $t \leq T$ <b>do</b>	
<b>if</b> $a \in \mathcal{A}_t$ <b>then</b>	
1. Find $R_a(k, t)$ for $\forall k \in \mathcal{K}_2$ and $B_\pi(a, t - 1)$ .	
2. Update the preference ordering of UAs and RBs, using (12), (13).	
3. Using $\succ_a$ , UA $a$ applies for the most preferred RB in $\mathcal{K}_a$ .	
4. Each RB $k$ accepts the most preferred UA, based on $\succ_k$ , among new applicants plus $\mu_t(k)$ , and rejects the rest. Next, $k$ is removed from applicants' $\mathcal{K}_a$ sets.	
5. Each UA $a$ calculates $\bar{R}_\pi(a, t)$ and updates $\succ_a$ .	
<b>repeat</b> steps 2 to 5 <b>until</b> $\mathbb{1}_a(\pi) = 0$ or $\mathcal{K}_a = \emptyset, \forall a \in \mathcal{A}_t$ .	
6. Data transmission occurs for each UA $a$ over $\mu_t(a)$ RBs.	
<b>else</b>	
I. UAs $a \in \bigcup_{t' > t} \mathcal{A}_{t'}$ send a request to SBS for mmW resource.	
II. SBS sets $J(t)$ to the number of received requests (from UAs with corresponding LoS UEs) and adjusts $\tau_1 = \tau_2/J(t)$ .	
III. SBS allocates mmW time-slots $j = 1, \dots, J(t)$ to applicants based on RR and updates $y_{ajt}$ variables.	
IV. If $y_{ajt} = 1$ , SBS transmits data at time-slot $j$ to UA $a$ . UA updates $B_\pi(a, t)$ .	
<b>end</b>	
$t \leftarrow t + 1$	
<b>end</b>	
<b>Output:</b> Stable scheduling policy $\pi^*$	

---

the proposed game can be classified as a *matching game with externalities*. For matching games with externalities, DA may not converge to a two-sided stable matching. Therefore, a new algorithm must be found to solve the problem.

To this end, we propose the context-aware scheduling algorithm shown in Table I. At each slot  $t$ , UAs apply for mmW and  $\mu$ W resources based on their local information. Steps 1-6 find the stable matching  $\mu_t$  for UAs that must be scheduled at  $\mu$ W band. Steps I-IV use RR to allocate mmW slots to UAs. For this algorithm, we have:

**Theorem 1:** The proposed algorithm in Table I yields a two-sided stable matching between UAs and  $\mu$ W RBs.

*Proof:* Since at any arbitrary time slot  $t$ , UAs  $a \in \mathcal{A}_t$  are involved in the matching game and  $\mathcal{A}_t \cap \mathcal{A}_{t'} = \emptyset, t \neq t'$ , it suffices to prove the two-sided stability of the matching algorithm at time slot  $t$ . The convergence of the algorithm in Table I at each slot is guaranteed, since a UA never applies for a certain RB twice. Hence, in the worst case, all UAs will apply for all RBs once, which yields  $\mathcal{K}_a = \emptyset, \forall a \in \mathcal{A}$ . Next, we show that once the algorithm converges, the resulting matching between UAs and RBs is two-sided stable. Assume that there exists a pair  $(a, k) \notin \mu_t$  that blocks  $\mu_t$ . Since the algorithm has converged, we can conclude that at least one of the following cases is true about  $a$ :  $\mathbb{1}_a = 0$ , or  $\mathcal{K}_a = \emptyset$ .

The first case,  $\mathbb{1}_a = 0$  implies that  $a$  does not need to add more RBs to  $\mu_t(a)$ . In addition,  $a$  would not replace any of  $k' \in \mu(a)$  with  $k$ , since  $k' \succ_a k$ . Otherwise,  $a$  would apply earlier for  $k$ . If  $a$  has applied for  $k$  and got rejected, this means  $\mu(k) \succ_k a$ , which contradicts  $(a, k)$  to be a blocking pair. Analogous to the first case,  $\mathcal{K}_a = \emptyset$  implies that  $a$  has got rejected by  $k$ , which means  $\mu(k) \succ_k a$  and  $(a, k)$  cannot be a blocking pair. This proves the theorem. ■

TABLE II  
SIMULATION PARAMETERS

Notation	Parameter	Value
$P_1, P_2$	Transmit power	1 W
$(\Omega_1, \Omega_2)$	Bandwidth	(1 GHz, 10 MHz)
$\omega_1, \omega_2$	Bandwidth per RB	480 KHz
$(\xi_1, \xi_2)$	Standard deviation of mmW path loss	(5.2, 10) [9]
$(\alpha_1, \alpha_2)$	Path loss exponent	(2, 3) [9]
$(\beta_1, \beta_2)$	Path loss at 1 m	(70, 38) dB
$\psi$	Antenna gain	18 dB
$\tau_2$	Time slot at $\mu$ W band	10 ms
$N_0$	Noise power	-174 dBm/Hz

#### IV. SIMULATION RESULTS

For simulations, we consider an area with diameter  $r = 500$  meters with the SBS located at the center. UEs are distributed uniformly within this area with a minimum distance of 5 meters from the SBS. Each UE has  $\kappa$  UAs chosen randomly and uniformly from  $T$  QoS classes. The main parameters are summarized in Table II. All statistical results are averaged over a large number of independent runs. We compare the performance with a “CSI-based” scheduler that relies only on the channel quality and disregards context information. The CSI-based scheduler assumes that all applications per UE can tolerate the same delay equal to the minimum of all corresponding UAs’ delays. We define the performance metric  $\lambda$  as the average number of QoS violations:

$$\lambda = \frac{1}{A} \sum_{t=1}^T \sum_{a \in \mathcal{A}_t} \mathbb{1}_a(\pi). \quad (14)$$

Due to the stochastic nature of wireless channels, we are interested in the statistics of  $\lambda$ , i.e., to evaluate whether  $P(\lambda \geq \lambda_{th}) \leq \epsilon$ , where  $\lambda_{th}$  is the maximum tolerable  $\lambda$  and  $\epsilon$  is a small probability. This can be written as  $F_\lambda(\lambda_t) \geq 1 - \epsilon$ , where  $F_\lambda(\cdot)$  is the cumulative distribution function (CDF) of  $\lambda$ .

Figs. 2(a) and 2(b) show the total amount of traffic transmitted at both frequency bands versus time for LoS probabilities of  $\rho = 0.5$  and  $\rho = 0.1$ , respectively. Moreover,  $T = 5$ ,  $M = 20$ ,  $\kappa = 5$ , and  $B = 0.5$  Mbits are considered. We observe that the traffic decreases over time, due to the finite buffer traffic model, as well as the fact that different classes of UAs are equally likely to be run at UEs. The result in Fig. 2 shows that the proposed algorithm exploits more mmW resources as more UEs are at LoS from SBS. In addition, the dual-band scheduling significantly increases traffic offload from  $\mu$ W band, reaching up to 43% at the fourth time slot.

Fig. 3 compares the performance of the proposed context-aware approach with CSI-based scheduling for  $M = 10$ ,  $\kappa = 3$ ,  $T = 5$ . Fig. 3 shows the average number of QoS violations  $\lambda$  as a function of the required load (number of bits  $B$  that must be transmitted per UA). In Fig. 3, we see that  $\lambda$  decreases with  $B$ , since more UAs meet their QoS constraint. Moreover, for  $\lambda_t = 0.01$ , Fig. 3 shows that the proposed approach can serve up to 60 kbits more traffic per UA compared to the CSI-based approach. For  $M = 5$  and  $\tau_2 = 10$  ms, this is equivalent to a 2.4 Mbps improvement in the average data rate. It worth to note that for  $\lambda < 0.01$ , the performance gain of the proposed approach against CSI-based approach remains unchanged. In addition, for

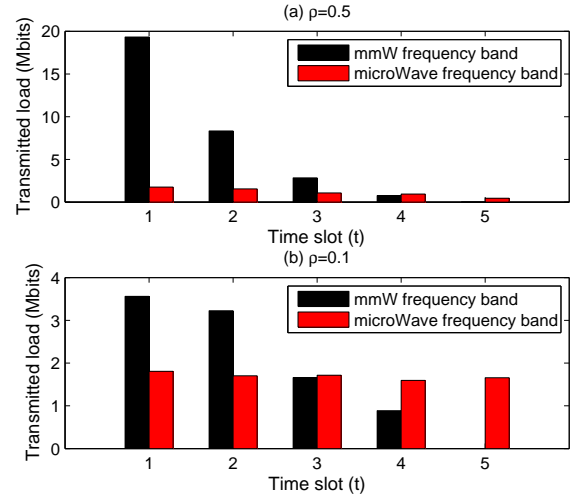


Fig. 2. Comparison of the transmitted load at each frequency band vs time slot, plotted for  $\rho = 0.5$  and  $\rho = 0.1$ .

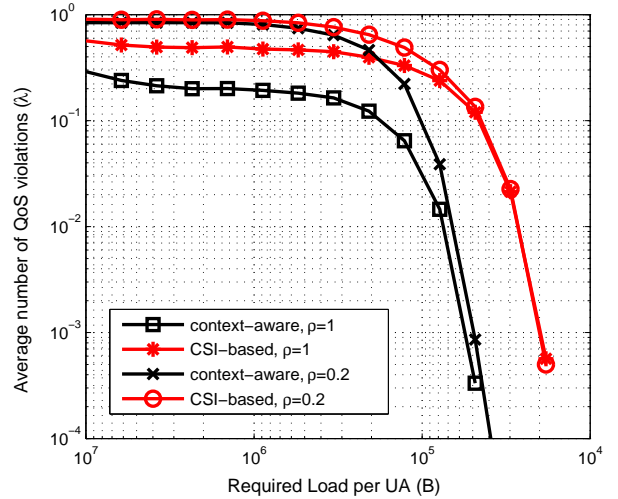


Fig. 3. Performance comparison of context-aware and CSI-based scheduling.

low loads, the result become independent of the probability of LoS, since UAs are scheduled in the  $\mu$ W band.

Fig. 4 shows the performance comparison in terms of the CDF of  $\lambda$ ,  $F_\lambda(\lambda)$ , for  $M = 20$ ,  $\kappa = 3$ ,  $T = 5$ ,  $B = 0.1$  Mbits, and  $\rho = 0$ . In Fig. 4(a), the 10 farthest UEs from the SBS are chosen as cell edge UEs and the statistics of  $\lambda$  for their corresponding UAs are shown. In contrast in Fig. 4(b), we choose the 10 UEs that are the nearest to the SBS as cell center UEs and the statistics of  $\lambda$  for their corresponding UAs are shown. With no LoS connection available for UEs, we observe that the performance of the cell edge UAs is poor for both approaches. However, Fig. 4(b) shows that the context-aware approach achieves more gain when cell center UAs are considered. For instance, using CSI-based approach, the probability of less than 10% of UAs be unsatisfied is 1%, while this probability is 37% for context-aware approach.

Fig. 5 shows the same performance metric as Fig. 4, with the same parameters except  $\rho = 0.3$ . Owing to available capacity at mmW band, the performance of both cell edge and cell center UAs are significantly improved compared to Fig. 4. Moreover,

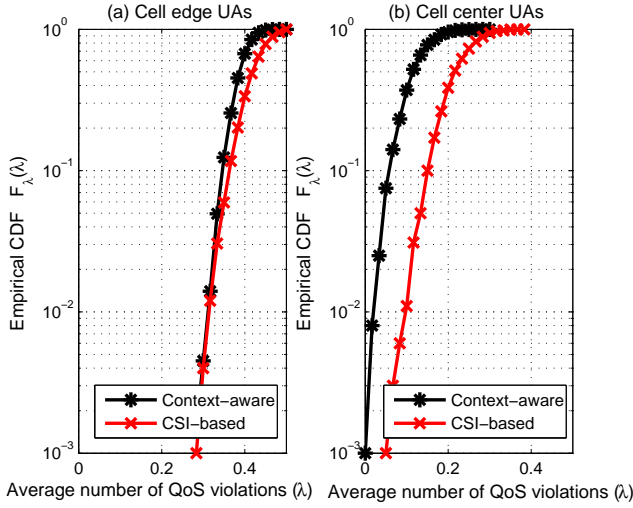


Fig. 4. Performance comparison between context-aware and CSI-based scheduling for  $\rho = 0$ , i.e., no LoS UE.

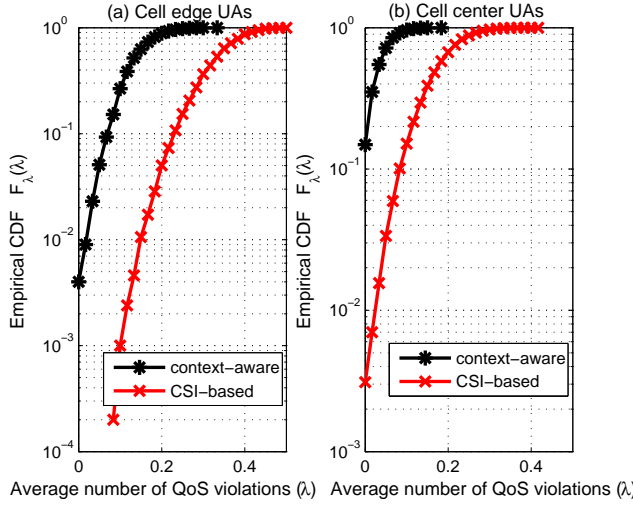


Fig. 5. Performance comparison between context-aware and CSI-based scheduling for  $\rho = 0.3$ .

we observe that the performance gap between the proposed context-aware approach with CSI-based approach increases as  $\rho$  increased. For instance, the proposed approach improves the QoS up to 90% (i.e. satisfying more UAs) for the cell edge UAs, for  $F_\lambda(\lambda) = 0.01$ .

The average number of iterations per time slot versus network size is shown in Fig. 6 for  $B = 0.5$  Mbits,  $M = 5$ , and  $\kappa = 3$ . The average number of iterations increases linearly with the number of UEs. In addition, we note that the average number of iterations decreases as  $\rho$  increases. This is due to the fact that more traffic can be offloaded to mmW band which enhances resource allocation at  $\mu$ W band. Overall, Fig. 6 shows that the proposed algorithm converges within a reasonable number of iterations even for large number of UEs.

## V. CONCLUSIONS

In this paper, we have proposed a novel context-aware scheduling framework for dual-mode small base stations operating at mmW and  $\mu$ W frequency bands. The proposed scheduler

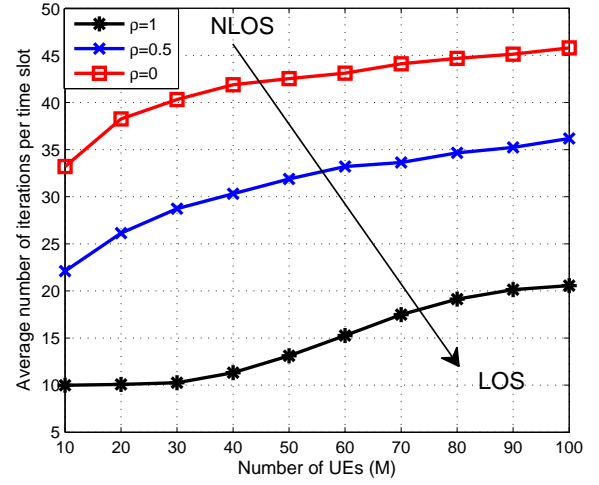


Fig. 6. Number of iterations for the proposed context-aware scheduling versus the number of UEs.

can provide delay guarantees per user application. We have formulated this context-aware scheduling problem as a one-to-many matching game which is then solved using a distributed algorithm. The proposed algorithm exploits mmW band resources for opportunistic traffic offloads, while guaranteeing the UAs' QoS. We have proved that the proposed algorithm yields a two-sided stable scheduling policy. Simulation results have shown the various merits and performance advantages of context-aware scheduling for dual-mode networks.

## REFERENCES

- [1] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, pp. 74–80, February 2014.
- [2] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.
- [3] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, pp. 3437–3458, October 2015.
- [4] H. Mehrpouyan, M. Matthaiou, R. Wang, G. Karagiannidis, and Y. Hua, "Hybrid millimeter-wave systems: a novel paradigm for hetnets," *IEEE Communications Magazine*, vol. 53, pp. 216–221, January 2015.
- [5] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Communications Magazine*, vol. 53, pp. 209–215, January 2015.
- [6] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, pp. 101–107, June 2011.
- [7] H. Singh, J. Oh, C. Kweon, X. Qin, H.-R. Shao, and C. Ngo, "A 60 GHz wireless network for enabling uncompressed video communication," *IEEE Communications Magazine*, vol. 46, pp. 71–78, December 2008.
- [8] D. Wu, J. Wang, Y. Cai, and M. Guizani, "Millimeter-wave multimedia communications: challenges, methodology, and applications," *IEEE Communications Magazine*, vol. 53, pp. 232–238, January 2015.
- [9] A. Ghosh, R. Ratasuk, P. Moorut, T. S. Rappaport, and S. Sun, "Millimeter-Wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152–1163, June 2014.
- [10] K. Seong, M. Mohseni, and J. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *IEEE International Symposium on Information Theory*, (Seattle, Washington), July 2006.
- [11] A. E. Roth and M. A. O. Sotomayor, *Two-sided matching: A study in game-theoretic modeling and analysis*. Cambridge University Press, 1992.
- [12] E. Jorswieck, "Stable matchings for resource allocation in wireless networks," in *Proc. of 17th International Conference on Digital Signal Processing (DSP)*, (Corfu, Greece), July 2011.